大数据时代的网络数据可视分析

时 磊[1] 童行行[2]
[1]中国科学院软件研究所
[2]美国亚利桑那州立大学

关键词：大数据 网络数据 可视分析

一图抵千言

清华大学 FIT 楼的环形长廊里，形式各异的科研宣传板上已悄然换上大数据相关的内容；私下抑或正式的研讨会上，“什么是大数据？”是讨论最多的内容。听过各种版本的诠释后，这个问题的答案似乎仍不明朗。但有一点却是肯定的，即大数据“难以采用现有关系型数据库存储”。本文将要探讨的——“网络（图）数据”，正是典型的带有这种特性的数据类型。近年来，国际顶级数据库会议 SIGMOD 和数据挖掘会议 SIGKDD 均有 3~5 个独立研讨会关注网络及图数据处理与分析。为使读者便于理解，本文采用“网络”来定义包含实体、实体间关联及其附加属性的一类数据集。与传统的网络研究相比，这里定义的网络更强调其作为数据的特征，而非作为互联网核心体系架构的角色。在数据分析相关领域，“网络”与“图”几乎可以等同视之。

在西方有一句格言：一图抵千言 (A picture is worth a thousand words)，意思是复杂的内容可通过一幅静态图像清晰地表达。这里强调了可视化作为人与数据之间纽带的作用。在大数据时代，这个作用愈发凸显，并以可视分析这种形式为人所知。今年春运期间，央视直播的百度迁徙图 [1] 是通过可视化形式直观分析海量数据的范例。在华丽的图表背后，汇聚了全国逾百个大中城市之间的上亿次人口流动，称得上是老百姓众包而成的大数据。巧合的是，这里展示的人口迁徙数据正是一种典型的网络数据，即以每座城市为实体信息，以城市间每次人口流动为实体间关联的网络数据集。类似百度迁徙图的做法在平民化推广可视化概念方面取得了一定程度的成功。然而，从技术角度看，大数据时代的网络（图）数据可视分析仍然存在极大的挑战。大数据的 5V 特征众所周知，即 Volume（体量大）、Variety（多样）、Velocity（快速）、Veracity（真实）、Value（价值）。本文试图从“大网络（图）数据”(Big Graph) 的 5V 特征出发，探讨大数据时代网络可视分析面临的新挑战。

5V特征的网络数据可视分析

网络规模：如何解开数据线团？

伴随着信息与互联网革命，我们生成并得以获取的网络数据体量已非经典的空手道俱乐部数据集 (Zachary’s Karate Club)（34 个节点）可比拟。国内四大微博注册用户超过 8 亿，每日新增微博约 2 亿条，用户与微博之间产生了巨大的关注、评论及转发网络；我们常用的 Windows/Linux 操作系统包含上千万行源代码，代码间存在百万次函数静态 / 动态调用、代码包含与继承关系；人类基因组工程揭示了生物体亿万细胞、蛋白质及基因之间的复杂交互关系。这些大容量网络的可视化形式在互联网上极少出现。采用常见算法来对上百万节点的网络进

行布局通常需要数小时或更长时间。即使对于仅有数千或几万节点的网络，大部分的可视化结果也非常杂乱，常被称为“线团”(hairball)。因此，如何利用可视分析手段，快速高效地解开数据线团，是本研究领域面临的首要挑战。

高维属性：如何展示附加数据?

大数据的多样性与近期网络可视化的研究热点——多维网络可视化 (multivariate network visualization) 不谋而合。这项研究重点关注的是节点及连接关系之上的高维属性。例如，著名的在线社会网络脸谱 (Facebook) 中，每个用户的档案包含个人信息、活动记录等近百个属性；从大量生物文献中利用文本挖掘技术抽取出的基因与蛋白质间的交互关系，分为刺激、抑制、中性等多种类型。传统的网络可视化方法通过节点（边）的大小、形状、颜色等视觉编码可呈现 5~10 维附加属性，但难以承载维数更高的新型网络数据。同时，这种编码方法不适于展现附加属性之间的潜在关联。如果将网络实体所涉及的多个网络整合起来，可形成体量更大、结构更复杂的异构网络，如在线用户真实身份的完整网络。这种节点与关系的混杂和不确定性为可视分析提出了更多的挑战。

动态特征：如何刻画网络变化?

我们所探讨的网络数据一直处于快速变化之中。人口流动网络随着日夜交替、春秋变换、时代演进而不断变迁，软件函数间的调用关系随着程序的执行进程而动态变化。如何利用可视化手段刻画网络的变化并分析总结其规律，是本领域研究者钻研多年的难题。虽然简单直观的“动画”式展示方法给用户以愉悦的使用体验，但多个重要研究结果表明，该方法难以适用于数据分析任务。用户往往过于关注可视化的动画效果，而忽略了网络数据的时序变化情况。其根结在于，网络节点在可视化过程中存在大幅度的位移，令用户失去分析目标。另一种常见的可视化方法是将时变网络的每个时间帧数据在单独的视图里展示。用户通过同时浏览多个时间帧的视图，比较、分析并记录网络的动态变化。这种方法未能充分利用大数据时代终端强大的分析计算能力，不适用于时间跨度较长、体量较大的动态网络数据。

真实展示：如何避免视觉骗局?

著名学者爱德华·塔夫特 (Edward R. Tufte) 早在 30 年前就已提出可视化谎言 (visualization lies) 的概念，即通过坐标轴变换、图形不等比缩放等易被忽视的手段，达到将数据内容刻意缩小 / 放大的视觉感知效果。这种做法过去常常在非专业场合如在线杂志、新闻的信息图中出现，而在当前的专业学术论文中已很少发生。但是，在大数据时代，由于网络数据本身存在来源、在线处理及整合等方面的误差，因此其不确定性的程度加大。同时在体量大、种类杂、变化快的网络大数据上实现可视分析，不可避免地涉及对数据的采样与抽象。如何保证用于可视分析的数据切片可代表整个网络数据集，以及如何避免其可视化表达引起用户对网络数据集的认知偏差，是本研究领域需要直面的问题。其突破口将不限于传统的可视化设计方法。

数据价值：如何深度服务用户?

毋庸置疑，网络数据可视分析的终极目标是为用户提供更高的价值。在与众多典型用户深入接触后发现，大部分用户使用网络可视化的理由是：“好看”“好玩”“给领导汇报”。虽然在软件工程及软件开发等少数领域，网络可视化已被认为不可替代，但在更多领域能否抓住大数据时代的契机，发掘出网络数据可视分析方法的核心价值，是本研究方向的关键挑战。曾有学者将网络可视化的用户任务划分为网络拓扑相关分析、节点（边）属性相关分析、浏览、宏观概要四类。在大数据时代，此用户任务分类是否仍然符合现状，能否定义出更贴近大数据分析核心功能的网络数据可视分析任务，都是亟待解决的问题。同时，可视化技术在领域应用也存在瓶颈，即难以为多种数据类型设计通用的可视化方法，因而需要代价昂贵的定制化工作。在大数据背

景下，在异构网络整合而成的统一类型的网络数据之上，能否突破定制化瓶颈，设计出一套完整的大数据网络可视分析平台，将决定网络可视分析未来的应用前景。

主流研究方向

深度网络分析与挖掘

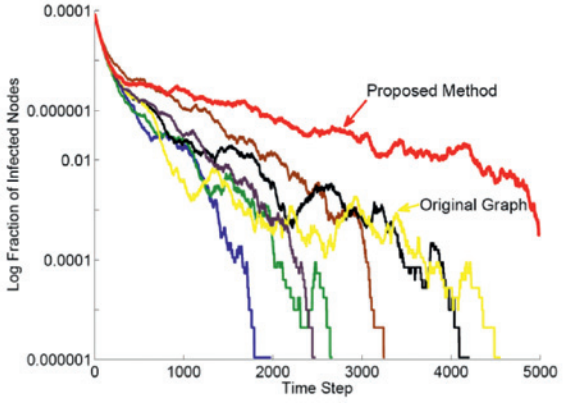
如何完成从“好看”、“好玩”到“好用”的角色转变，是大数据时代网络可视分析所要解决的首要问题。从数据分析的角度来看，我们需要从体量大、多变、多样的网络数据中发掘出深层的知识或模式。比如，给定一组网络数据，我们可从以下三个层面进行分析挖掘。(1) 在全图层面上，分析网络的一些全局统计量，比如网络节点度的分布、网络的直径、网络谱（特征值）的分布等。(2) 在子图层面上，可对原图作适当的分割，以发现全图所含的聚类结构。(3) 在节点层面上，刻画不同节点之间的相似度、相关性甚至因果关系。例如在文献 [2] 中，我们提出了一种新型分析方法，用以定位大型网络中信息传播的关键链接，如图 1(a) 所示。若将这些结果以适当的方式呈现给终端用户 [3]，则可以极大地提高网络数据可视分析方法的有效性，如图 1(b) 所示。

任务驱动数据变换

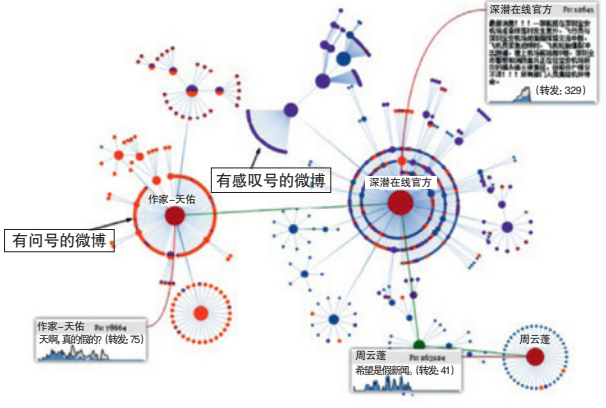
针对海量数据，美国马里兰大学本·施耐德曼 (Ben Shneiderman) 教授曾提出“纵览为先，缩放并过滤，按需查看细节 (Overview first, zoom and filter, then details on demand)”的以可视化方式查询信息的黄金法则。由此可看出，纵览（或称为可视化摘要）在可视分析流程中的重要性。然而，随着大数据 5V 特性的扩散，从数据全局提取摘要面临巨大的计算开销，同时对大数据的所有特征进行摘要可视化也变得极为困难。一系列新方法试图根据用户的特定任务，采用定制的数据变换生成大型网络的摘要可视化。文献 [4] 提出了个人中心动态网络的可视化方法，如图 2(a) 所示，可用于在短信网络中用可视化方式检测垃圾短信发送者与正常使用用户的不同行为特征。文献 [5] 针对单篇学术论文主题发展的分析任务，定义了影响力流动最大化的可视化视图生成标准，并采用基于矩阵分解的算法构建可视化摘要，如图 2(b) 所示。上述两种方法均不需要处理整个网络，仅须对特定节点的子网进行有针对性的数据变换。

自然和谐用户交互

人机交互方法是可视化界面不可或缺的组成部分。典型的网络数据可视化交互，如视图缩放、平移、节点拖拽，已得到大量应用并随着众多的开源

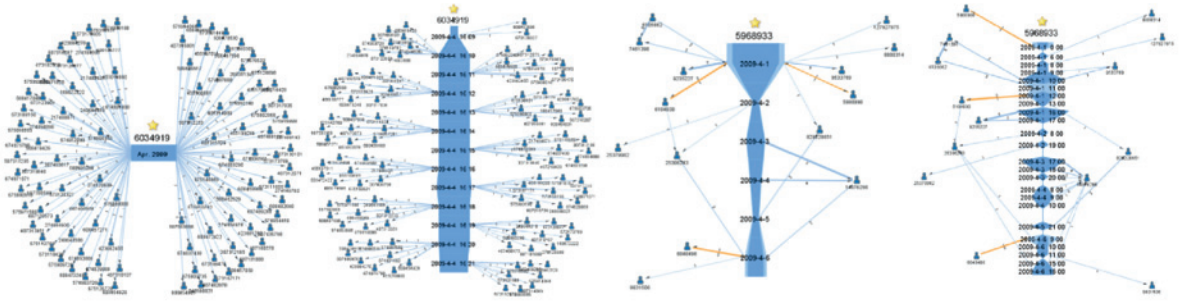


(a)边操控方法性能[2]



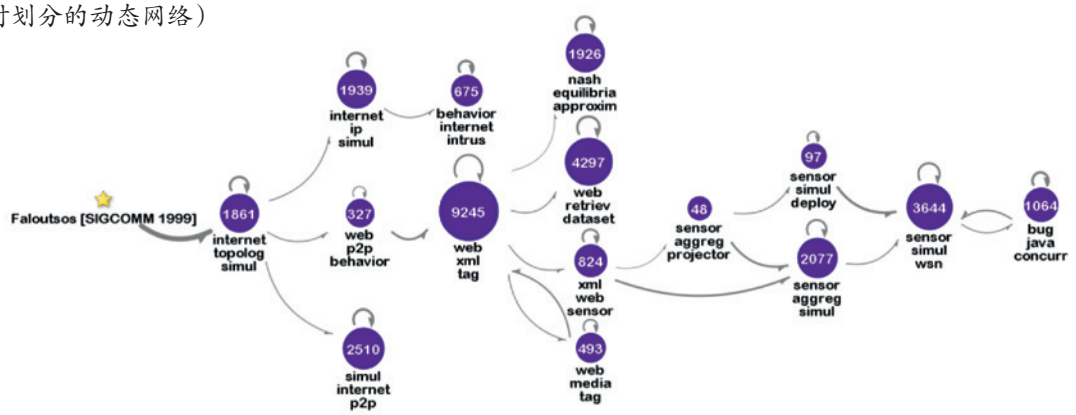
(b)微博系统中采用的微博信息传播可视化方法[3]

图1　大型网络中信息传播的深度分析及可视化方法



(a) 1.5维个人中心动态网络可视化方法[4]

（数据集自左至右依次为典型垃圾短信发送者按月、按天划分的动态网络，典型正常使用用户按天、按小时划分的动态网络）



(b) 基于影响力流动最大化标准生成的单篇论文影响图谱[5]

（数据集为ArnetMiner v6学术引用网络，初始节点为Faloutsos三兄弟1999年发表在*SIGCOMM*上关于互联网拓扑幂定律的经典论文）

图2 基于特定数据变换的网络可视化方法[6]

图3 基于层次聚类的大型网络可视化系统HiMap

（自左至右为执行层次递进操作后的逐帧连贯动画过程。层次递进后的顶层网络为左1图中蓝色轮廓线标注的子网）

网络可视化工具包广泛传播。在大数据时代，网络数据的可视分析至少在三个方面亟须创新的交互方法设计。(1) 网络浏览。如前所述，大数据可视化多从全局纵览视图出发，但最终停留于关键细节上的知识或模式发现。用户交互浏览方法正是连接全局视图与细节发现的重要桥梁。与传统方法不同的是，此类网络数据交互常涉及网络层次可视分析，如层次递进与遍历。图 3 展示了文献 [6] 提出的大型网络可视化系统 HiMap 的层次递进交互方法。(2) 网络比较。在大型网络尤其是动态网络的分析过程中，多个子网或者不同时间帧的网络比较是一类典型的用户任务。在网络可视化中集成专用的交

互比较方法可极大地提升用户执行该任务的能力。(3) 算法（参数）调优。大数据集的特性导致其无法仅仅依靠单一算法的运用或通过预设进行参数分析，而算法及参数调优通常是分析人员时间开销最大的环节。采用可视化界面来直观展现算法调优过程的动态变化，将有助于改进大数据分析的效率。

多态网络可视隐喻

当前互联网及学术期刊 / 会议上绝大多数的网络数据采用节点 - 边的传统可视化隐喻。这与节点 - 边在形式上符合人类对于关联数据的感知规律有关，详细讨论可参阅格式塔理论 (Gestalt theory)。在大数据时代，网络的体量、动态性及丰富的节点 / 边属性给这一传统带来了挑战。研究结果表明，对于较大型的网络，基于邻接矩阵形式的可视化在多个典型网络分析任务中显著优于节点 - 边的可视化形式。文献 [7] 进一步提出了混合节点 - 边与邻接矩阵表达的可视化方法 NodeTrix（如图 4）。针对动态网络数据，基于平行坐标隐喻的平行边分割方

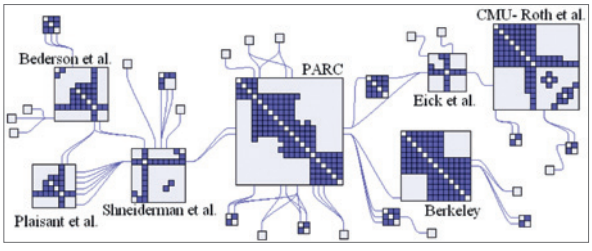


图4 采用网络/矩阵混合方式(NodeTrix)展示信息可视化领域的主要学术合作关系网络[7]

法可提供更佳的可扩展性，以展示大型网络。带有文本信息的网络数据可通过标准的字云 (word cloud) 可视化与节点 - 边表达的结合，展示网络与文本混合的异构数据。

结语

网络数据蕴含世间万物相联、相生、相克的复杂关系。在大数据时代，网络数据正走向大型化、动态化、多样化。如何通过可视化分析方法诠释数据特征，真实展现网络原貌，最终深度服务用户，是信息可视化、数据挖掘及人机交互领域面对的共同挑战。本文基于大数据 5V 特征所引出的技术问题，总结了当前学术界正在实践的四条研究路线。需要说明的是，关于大数据时代的网络数据可视分析问题，学术界及工业界目前尚无系统、完整的方案与方法论。我们期待在未来十年或更长时间内，该问题能得到最终解决。这需要整个研究领域的持续投入与共同协作。■

时 磊

CCF会员。中国科学院软件研究所副研究员。主要研究方向为信息可视化、可视分析、数据挖掘。shijim@gmail.com

童行行

美国亚利桑那州立大学助理教授。主要研究方向为大数据挖掘。
hanghang.tong@gmail.com

参考文献

[1] http://qianxi.baidu.com/.

[2] H.Tong, et al., Gelling and Melting. Large Graphs through Edge Manipulation, *ACM CIKM*, 2012: 245~254.

[3] D. Ren, X. Zhang, Z. Wang,et al... WeiboEvents: a crowd sourcing Weibo visual analytic system. *IEEE Pacific Visualization Symposium*, 2014: 330~334.

[4] L. Shi, C.Wang and Z.Wen. Dynamic network visualization in 1.5D. *IEEE Pacific Visualization Symposium*, 2011:179~186.

[5] L. Shi, H. Tong, J. Tang and C. Lin. Flow-based influence graph visual summarization. http://arxiv.org/abs/1408.2401.

[6] L.Shi, et al.. HiMap: adaptive visualization of large-scale online social networks. *IEEE Pacific Visualization Symposium*, 2009: 41~48.

[7] N.Henry, J.D.Fekete and M.McGuffin. NodeTrix: a hybrid visualization of social networks. *IEEE Transactions on Visualization and Computer Graphics*, 2007, 13(6): 1302~1309.